Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Name: Soma Pavan Kumar
Email: spkumar1998@gmail.com
Contribution:

1. Data Importing and Cleaning
2. Data Visualizations and getting insights
3. Getting the Statistics of Data
4. Data Modeling with various machine Learning algorithms
    1. Logistic Regression
    2. K-Nearest Neighbors
    3. Naive Bayes Classifier (Gaussian)

5.Conclusion

Please paste the GitHub Repo link.

Github Link: -

https://github.com/PavanKumar181098/Company-Bankruptcy-Prediction

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Write here the short summary

Prediction of bankruptcy is a phenomenon of increasing interest to firms who stand to lose money because of unpaid debts. Since computers can store huge data sets pertaining to bankruptcy, making accurate predictions from them beforehand is becoming important.

The data were collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange.

In this project, we have used various classification algorithms like Logistic Regression, Knn and Naive Bayes(Gaussian) Classifiers and their accuracy metrics have been plotted simultaneously.

For the First step, the data is imported from a csv file and converted to a panda's DataFrame. Pandas is an inbuilt library in python that is used in handling and manipulating Dataframes. Dataframe is basically a collection of Rows and Columns.

Once the data is clean, we check for the statistics of the cleaned data. The statistics of the data tells us the mean, median and the distribution of the data and some more info. Also, we checked for any null values and duplicate values in the whole dataset and for our surprise, we didn't find any null nor duplicate values.

After the data is cleaned, Visualizations can be done on the data and inferences can be noted.

One such visualization is to find the correlation between the variables. Plotting the correlation between variables is not visually appealing as there are 65 different variables. But we see some red squared in between that says they are highly correlated.

As the target variable is categorical i.e., Bankrupt or not, plotting this variable we find out that around 97% of the data present is of financially stable and the rest is financially unstable.

Similarly, we can plot the Boxplot of Correlations between the variables and also feature distributions for close to bankruptcy companies.

After all this, we remove the outliers. Outlier is a data point that is present very far away from the mean of the distribution of the data. There is a set procedure to remove the outliers. We plot the graphs of some of the features to take a loot at the graphs after removing the outliers.

Then, we perform a Log (Logarithmic) transformation on the data, the log transformation is performed in the data to remove any skewness present in the data and produce an approximate log-normal distribution. After the log transformation has been performed, we see an approximation of normal distributions in our dataset.

Coming to the second part, which is Data Modeling in which we try out different Supervised Machine Learning Classification algorithms fits the best to our dataset and gives us the best accuracy scores like precision, Recall, AOC-ROC curve, Accuracy and so on

Starting with a primitive Logistic Regression, Logistic regression is a statistical model that in its basic form uses a logistic Function to model a binary dependent variable, although many more complex extensions exist. After we apply Logistic regression to our model the AOC-ROC score of 0.84

Next, we take the K-Nearest Neighbors Classifier algorithm. KNN being a much more complex algorithm produces better results. Providing an overall accuracy of 0.897, f1-score of 0.40 and finally AOC-ROC score of 0.625

Let's look at the next Algorithm Naive Bayes Classifier(Gaussian). Naive Bayes is an implementation of gradient boosted decision trees designed for speed and performance. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Naive Bayes classifier is also used for text classification algorithms. Coming to its accuracy scores, overall accuracy score of 0.66, and a f1-score of 0.16 and AOC score of 0.76

Concluding, according to the f1-scores and the Accuracy metrics, KNN is the best model to choose for further predictions. According to AOC-ROC, Logistic Regression stands first with a score of 0.86

Nevertheless, in this case, the best decision is to use Logistic regression because it can better recognize the minority class even misclassifying some not close to bankruptcy companies as close to bankruptcy.

Also, the parameters that contribute to bankruptcy of a company are also external such as decisions taken by the CEO of the company and dwindling workforce and so on, which can neither be calculated nor predicted.